reinforcement learning: planning and control through experience

Gavin Taylor

United States Naval Academy

Annapolis, MD 21401

Abstract

Significant recent advances in Machine Learning have resulted in an explosion of interest in the field and its possible applications. Reinforcement Learning is a sub-field of ML which attempts to address the control question, including in the control of difficult-to-model, dynamic systems. Reinforcement Learning has undergone particular growth, with several recent high-profile successes. In this paper, I introduce the field of Reinforcement Learning, and describe its strengths and weaknesses to help the reader understand if the approach is appropriate for their control problem.

Keywords

Reinforcement Learning, Control Theory, Machine Learning, Artificial Intelligence.

Introduction

The layout and style to be adopted in the preparation of your paper are described and illustrated through an example. ***Authors are requested to strictly adhere to these instructions***. If you are already familiar with Word, then the FOCAPD 2019 template should be of great help for you since it will change the layout so as to suit FOCAPD 2019 requirements. Nonetheless, the use of the document template is not mandatory.

Authors are expected to submit their paper by   
December 1, 2018. Only PDF files will be accepted for upload. The papers should be prepared if possible using Microsoft Word. LaTex can be used provided the format is similar to the one of this template.

As an author, you are fully responsible for the quality and appearance of your contribution. We request that you observe carefully these instructions for preparing your manuscript. *In almost all situations, use of the built-in template styles will save you from the grim details of fonts, sizes, line spacings, indentation, etc.* Failure to produce a paper in proper format will involve: (1) a request for a reformatting to meet the standards, or (2) the possibility that your contribution will not appear in the final Conference Proceedings.

Problem Statement

Imagine a child learning to ride a bicycle. From their senses, the child is confronted with a continued stream of useful information that describes the current state of their bicycle-riding adventure. For example, this data includes the angle of their handlebars, the amount of weight being shifted from one side of the bike to the other, the speed at which the bicycle is being ridden, an observation of the type of surface being ridden on, and any pain in their knee there may be from previous falls and scrapes. After many experiences and experiments, the child has sufficient data to learn to ride, maximizing the positive parts of the experience (euphoria from speed and independence, parental congratulations), while minimizing the negative aspects (falls, scrapes, embarrassment).

Mathematically, we attempt to describe this problem with a Markov Decision Process (MDP). An MDP is a tuple , where is the set of all states the problem might be in (in the case of a bicycle, this could include angle from upright, the angle of the handlebars, the current speed, the road surface, etc.); is the set of all actions the agent can take (turn the handlebars, pedal at some speed, etc.), is a transition kernel describing the probabilities of transitioning between two states when performing an action (, where ); is a reward function (for example, positive numbers for states in which the rider has reached a goal, or negative numbers for states in which the rider’s knee hurts), and is a discount factor, which is used to describe how much we prefer rewards in the short term to rewards in the long term.

The goal of the agent moving through the states of this MDP is to learn behavior (mathematically, a policy ) which maximizes the rewards received, as expressed by the Bellman equation, . This policy is learned from a number of samples of the form . A wide variety of approaches of learning effective policies exist, and have been active areas of research for decades.

Progress and Successes

The Bellman equation was introduced in the late 1950s. Progress in the decades since are best reviewed in Sutton and Barto (2018). Common experiments included some very basic control problems, including the Inverted Pendulum, in which a rod on a hinge must be kept balanced upright by moving its base left and right, and the Mountain Car, when an underpowered car in a valley must gain momentum by accelerating and reversing.

Very recently, however, neural network theory and GPU hardware have combined to produce some extremely effective agents on much more difficult problems. In 2015, Mnih et al. published a paper in which deep neural networks approximated the Bellman equation accurately enough to produce a policy capable of playing many Atari games at a super-human level. In the few years since, deep neural networks have produced agents capable of performing a wide range of difficult planning and control problems, including robotic control, humanoid mobility, and the world’s best players of the games of Go and chess (for example, Shulman et al., 2015 and 2017; Mnih et al., 2016; and Silver et al., 2017 and 2018), all purely from experience and without any human modeling or knowledge.

This research has resulted in a number of easily-deployed algorithms that can learn useful policies on a wide variety of tasks. This opens the door for the application of RL to real-world control problems.

Advantages

The Bellman equation was introduced in the late 195

Disadvantages

The Bellman equation was introduced in the late 195

Acknowledgments

This section should be placed at the end of the paper before the **References** section. This section should include acknowledgments of financial, institutional and personal support.

References

References (*Reference* style) should be listed in alphabetic order by surname at the end of the paper following the section title **References** right after the Acknowledgment section. Do not begin a new page with the Reference section unless it is absolutely necessary. The References section, with the exception of the title, should be typed in 9-point font. Each line of a reference, except the first, should be indented 0.5” from the left margin. All references included in the list should be cited in the text and vice versa. References are to be cited in the text by author’s name (year), or (authors’ names, year). For instance, Smith and Jones (1970), or (Ekeland et al., 1985). The expression et al. should be used when there are three or more authors. Two or more papers by the same authors published in different years should be chronologically arranged. Two or more papers by the same authors published in the same year should be distinguished by letters after the year.

Mnih, V., et al. “Human-level control through deep reinforcement learning.” *Nature* 518.7540 (2015): 529.

Mnih, V., et al. “Asynchronous methods for deep reinforcement learning.” *International conference on machine learning*. 2016.

Shulman, J. et al. “Trust region policy optimization.” *International conference on machine learning*. 2015.

Shulman, J. et al. “Proximal policy optimization algorithms.” *arXiv preprint arXiv:1707.06347* (2017).

Silver, D., et al. “Mastering the game of Go without human knowledge.” *Nature* 550.7676 (2017): 354.

Silver, D., et al. “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play.” *Science* 362.6419 (2018): 1140-1144.

Sutton, R. S., Barto, A. G. (2018). Reinforcement learning, An introduction. *MIT Press*.